Comparing and validating hypothesis test procedures: graphical and numerical tools

Pedro Delicado and Iolanda Placencia

Departament d'Economia i Empresa, Universitat Pompeu Fabra

Ramon Trias Fargas 25-27, 08005 Barcelona, SPAIN

April 28, 1997

Abstract

When the behaviour of a specific hypothesis test statistic is studied by a Monte Carlo experiment, the usual way to describe its quality is by giving the empirical level of the test. As an alternative to this procedure, we use the empirical distribution of the obtained p-values and exploit its information both graphically and numerically.

Keywords and phrases. Simulation, graphics, goodness of fit, distances between distribution functions.

1 Introduction

In a hypothesis test problem several statistics can be employed to test the same null hypothesis. The theoretical distribution of these test statistics is unfortunately unknown in many cases. As a consequence, it must be properly approximated and usually there is not a unique way to do it. So each test statistic and each specific approximation to its distribution may lead to a different p-value for the same test and possibly to a different decision about H_0. To discern between proposals simulation methods are a highly useful resort. The scope of this paper is to process the output of those Monte Carlo experiments graphically and numerically beyond the conventional practice of reporting empirical levels.

In essence, a Monte Carlo experiment lies in simulating a number of S samples of size n. From each sample a different value of the test statistic T_n and its corresponding p-value, computed from the distribution of T_n or from an approximation, are registered. The empirical significance level of the test, $\hat{\alpha}$, is traditionally reported as the only reference to decide if the test strategy is better than other schemes of inference. This conventional method focuses attention on a single value of the significance level α. It does not take into account the behavior of the statistic at other levels of significance. Consistently, to provide only such a specific reference can hide valuable information about the whole behavior of the statistic.

For that reason, in this work we make use of the information contained in all the p-values obtained by simulation. The empirical distribution function of the simulated p-values encloses all the relevant aspects of the Monte Carlo experiment, and it can be exploited graphically and numerically in terms of the nearness of that empirical distribution to the distribution of a uniform random variable in $[0, 1]$.

The paper is organized as follows: section 2 introduces notation and describes previous related works; definitions of several distances between distribution functions, multivariate analysis of simulated distances and the tabulation of some of them are the scope of section 3; section 4 contains two practical examples: first, the test of the lack of correlation, and second, the test of the equality of variances for two independent samples. In the second example, some considerations about reflecting the power of tests are expounded. Finally, in section 5, we draw a few conclusions.

2 Some notation and concepts

Consider the hypothesis test H_0 against H_1, concerning some aspect of the distribution of a random variable X. Given a sample $X_1, \ldots, X_n$ from X, the p-value is the smallest value of the significance level α at which the null hypothesis is rejected. For the clarity of the exposition we assume that the critical region can be written as $R_\alpha = \{(X_1, \ldots, X_n) \in \mathcal{X}^n : T_n(X_1, \ldots, X_n) \geq F_{T_n}^{-1}(1-\alpha)\}$, where F_{T_n} is the distribution of the test statistic T_n under the null hypothesis. First, we consider this distribution is known. In that case, the p-value is a random variable that takes the value $p\text{-value} = p\text{-value}(X_1, \ldots, X_n) = 1 - F_{T_n}(T_n(X_1, \ldots, X_n))$. This random variable is uniform in the interval $[0, 1]$ under H_0. The decision of rejecting or

not the null hypothesis at a given nominal size α can be expressed in terms of the p-value: the null hypothesis is rejected if and only if the p-value is lower than the significance level α.

Sometimes the exact distribution F_{T_n} of T_n is unknown. Then, to make inference, it is necessary to approximate F_{T_n} in some way. If an asymptotic approximation to F_{T_n} is available, it is usually used to define the p-value of the test. There exist situations where the asymptotic distribution of T_n is also unknown. In some of these cases, bootstrap techniques are used to approximate the distribution of T_n. This happens when a limit distribution exists for the statistic T_n and a bootstrap version of T_n (e.g., T_n^*) has the same limit distribution. The approximation of F_{T_n} is the the empirical distribution of a sequence of B bootstrap observations of T_n^* (see Efron and Tibshirani 1993 for an introduction to bootstrap techniques).

When some approximation $\tilde{F}_{T_n}$ (obtained from either asymptotic arguments or bootstrap) is involved in the process, a Monte Carlo experiment is worthwhile in order to investigate whether it is an acceptable estimate of the theoretical distribution F_{T_n}. If the S simulated samples are according to the null hypothesis and F_{T_n} is known, the S obtained p-values form a sample of a $U([0,1])$ distribution. For a given theoretical significance level α, the empirical level of the test is $\hat{\alpha} = \frac{1}{S}\sum_{s=1}^{S} I_{(0,\alpha]}(p_s) = \hat{F}_p(\alpha)$, where p_s is the p-value in the s-th replication of the experiment and $\hat{F}_p$ is the empirical distribution of the sample $\{p_s : s = 1, \ldots, S\}$. The random variable $\hat{\alpha}$ follows a distribution $B(S,\alpha)/S$. In case that F_{T_n} is approximated by some distribution function $\tilde{F}_{T_n}$, then the distribution of the p-values and the empirical level $\hat{\alpha}$ are only approximately $U([0,1])$ and $B(S,\alpha)/S$ distributed, respectively.

A graphic of the empirical function $\hat{F}_p$ permits to appreciate if it is near the unity square diagonal (the distribution function F_U of a $U([0,1])$ random variable). In that case, we deduce that $\tilde{F}_{T_n}$ is near F_{T_n}. The use of this kind of graphics in the literature is not new. Delicado (1995) develops some goodness of fit tests for the distribution of the coefficients in the random coefficient regression model, and reports the results of simulation experiments using graphics of the empirical distributions $\hat{F}_p$. Their usefulness is proved there when data are generated from both the null and the alternative hypotheses. Davidson and MacKinnon (1994) discuss the graphic of $\hat{F}_p$ (they call it *p-value plot*) and propose a variation, the *p-value discrepancy plot*, consisting in plotting the pairs $(p_s, \hat{F}_p(p_s) - p_s)$, where s index the simulated samples. The last graphic is more appropriate when two different tests with

good behavior under H_0 have to be compared. Authors also refer to other published papers where the empirical distribution function $\hat{F}_p$ or its inverse (the quantile function) have been used as a natural way to summarize the results of simulation studies. They are particular cases of the well known P-P and Q-Q plots (see, for instance, Chambers et al. 1983).

The main advantage of using p-value plots instead of reporting only some empirical values is that plots ensure us that the performance of a test is adequate at all the significance levels. In particular, when we report a p-value as the result of a certain hypothesis test, we must be sure that the test procedure has an empirical level similar to the theoretical one at nominal sizes near to the p-value we pretend to report, not just for two or three nominal sizes previously chosen.

Not only graphics can be derived from $\hat{F}_p$. We have already mentioned that empirical levels can be calculated from $\hat{F}_p$. In addition to them, many numerical single values that measure the nearness of $\tilde{F}_{T_n}$ to F_{T_n} can be extracted from $\hat{F}_p$. Distances between distribution functions computed from $\hat{F}_p$ and F_U gives us very rich insights to know whether a testing procedure is correct or not, much better than the empirical significance levels do. The use of Kolmogorov-Smirnov distance was presented in both Delicado (1995) and Davidson and MacKinnon (1994).

An important drawback is inherent in p-value plots. Davidson and MacKinnon (1994) word it as follows: *because they use one dimension for nominal size, p-value plots and p-value discrepancy plots cannot use that dimension to represent something else, such as the value of some parameter.* For example, to represent the behavior of a test procedure against some alternative hypotheses, we need not a single plot, but several plots. Distances between $\hat{F}_p$ and F_U solve this problem and continue to exhibit a global view through the whole range of possible p-values.

3 Distances between distribution functions

The graphic study of nearness between $\hat{F}_p$ and F_U by means of p-value plots is strengthened when we compute some distances between these distribution functions. The ones proposed here are based on the Kolmogorov-Smirnov distance and on L_1 and L_2 norms. L_1 is known as Mallows distance and it is related to the Gini's index used in Economics, and L_2 is the squared root of the Cramér-von Mises distance (see, for instance, Shorack and Wellner 1986,

$$d_\alpha = |\hat{\alpha} - \alpha|, \alpha \in (0,1)$$

$$d_{KS} = \sup_{p\in[0,1]} |\hat{F}_p^S(p) - p| \qquad d_{KS}^w = \sup_{p\in[0,1]} |\hat{F}_p^S(p) - p| w(p)$$

$$d_{L_r} = \left(\int_0^1 |\hat{F}_p^S(p) - p|^r dp\right)^{1/r} \qquad d_{L_r}^w = \left(\int_0^1 |\hat{F}_p^S(p) - p|^r w(p) dp\right)^{1/r}$$

Table I: Some distances between the empirical distribution function of p-values and F_U.

chapter 3.8, for these and other definitions of distances between distribution functions). As we are dealing with hypothesis tests, distances should be more sensitive to deviations of $\hat{F}_p$ from the diagonal at low values of the nominal size α. The inclusion of a weight function $w : [0,1] \to \mathbb{R}^+$ in the distance definitions allows us to pay more attention to that rank of values.

The distances we have considered are displayed in Table I. In their definitions, $w(p)$ is a weight function defined in $[0,1]$ with integral equal to one, that takes higher values at low p's. These distances are very jointly related because, in fact, all of them measure discrepancies between two distribution functions.

Under the null hypothesis, the theoretical distribution of p-values is F_U. Thus, the standardized empirical process $\nu_S = \sqrt{S}(\hat{F}_p^S - F_U)$ converges to the standard brownian bridge as S goes to infinity, where S is the number of simulated samples. Moreover, the distances we have defined above and many others, standardized by the factor $\sqrt{S}$, can be expressed as continuous functionals of the empirical process ν_S. So if we consider k distances $d_1, \ldots, d_k$, the random vector $(d_1, \ldots, d_k)'$ is a k-dimensional continuous functional of ν_S and, by the Continuous Mapping Theorem, has a limit joint distribution that is the same functional applied to the brownian bridge. Theory about that kind of properties can be found, for instance, in Shorack and Wellner (1986). The limit distributions of the unweighted distances we have introduced are explicitly derived in that book, and they are also tabulated.

The distances we propose are continuous functionals of ν_S and they generally have complicated unidimensional distributions. Moreover, when some weight function is introduced, the standard theory cannot be applied directly and it is not possible to find published tables of the resultant distances. At our knowledge, it does not exist neither any study referred to the joint distribution of several distances. Monte Carlo methods are useful to know about this joint distribution.

3.1 Multivariate analysis of simulated distances

We present a simulation study where observations of the joint distribution of nine distances between $\hat{F}_p^S$ and F_U are generated. Sample sizes $n = 30$, 50, 100, 500 and 1000 are used and 10000 samples are drawn for each case.

Nine standardized distances are analyzed: $d_i = \sqrt{S} d_{\alpha_i}$, $i = 1, 2, 3$, where $\alpha_1 = .01$, $\alpha_2 = .05$, and $\alpha_3 = .1$, $d_4 = \sqrt{S} d_{KS}$, $d_5 = \sqrt{S} d_{KS}^w$, $d_6 = \sqrt{S} d_{L_1}$, $d_7 = \sqrt{S} d_{L_1}^w$, $d_8 = \sqrt{S} d_{L_2}$ and $d_9 = \sqrt{S} d_{L_2}^w$. Table I shows definitions of these basic distances. The density function of a random variable with distribution $\beta(a = 2, b = 8)$ as w has been chosen here. Thus,

$$w(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} = 72\ p(1-p)^7, \quad 0 \le p \le 1.$$

Our objective is to discover the relationships between the distances $d_1, \ldots, d_9$ and to see if they are reporting similar or different information about the nearness between $\hat{F}_p$ and F_U.

We carry out a multivariate analysis of the data matrix $D_{10000\times 9}$ that contains in each row the observed distances $(d_1, \ldots, d_9)$ between F_U and the empirical distribution functions of simulated samples from $U([0,1])$. We present results only for $n = 1000$. For other sample sizes, conclusions are similar but less clean.

A quick look at the values of single columns of D shows that d_i's are very asymmetric. So we decide to transform D by taking logs on d_i. Let $\tilde{D}$ be the transformed data matrix. The computed sample correlation matrix of data $\tilde{D}$ is

$$\begin{bmatrix}
.12 & & & & & & & \\
.04 & .29 & & & & & & \\
.01 & .09 & .19 & & & & & \\
.03 & .27 & .56 & .44 & & & & \\
.03 & .14 & .26 & .90 & .52 & & & \\
.06 & .29 & .52 & .62 & .88 & .73 & & \\
.02 & .11 & .24 & .94 & .50 & .99 & .71 & \\
.04 & .25 & .47 & .68 & .85 & .77 & .99 & .75
\end{bmatrix}$$

A Principal Component Analysis is done on the correlation matrix of $\tilde{D}$. The percentages of variance explained by the first components are shown in Table II.

Figure 1 shows the distances $\tilde{d}_i$ placed on the plane of the first two components. The first component is a weighted mean of distances, in which

Comp.	Percentage of variance	Cumulative Percentage
1	56.18	56.18
2	15.83	72.01
3	11.19	83.20
4	8.55	91.75
5	5.65	97.40
⋮	⋮	⋮

Table II: Percentage of variance explained by the first principal component of $\tilde{D}$.

distances based on empirical sizes have lower weight than the rest. This first component can be interpreted as an index of discrepancy between each simulated sample and the uniform distribution. Global distances have the most important role on defining this index.

In Figure 1 we also can see the relative position of each $\tilde{d}_i$ with respect to the others. They are placed on an imaginary half circle, and if we go through it, we first find distances based on empirical values ($\tilde{d}_1$, $\tilde{d}_2$ and $\tilde{d}_3$, in that order), then we arrive to weighted distances ($\tilde{d}_5$ followed by $\tilde{d}_7$ and $\tilde{d}_9$) and finally to unweighted distances (quite near one another). This relative position confirms that the behaviour of the test procedure at small values of nominal sizes, gives not many clues about its global behaviour. The information that unweighted distances are offering is very different to that reported by empirical values. Weighted distances are a compromise between the two other groups.

Third, fourth and fifth components are closely related to $\tilde{d}_1$, $\tilde{d}_2$ and $\tilde{d}_3$, respectively. Therefore we can deduce that those distances present particularities that do not concern the global goodness of fit of the p-values to the uniform distribution.

We conclude that a weighted distance is the best choice if we want to use only one measure of nearness from p-values to F_U. The choice of d_7 or d_9 (the weighted L_1 or L_2 norms) seems to be the most reasonable.

A similar analysis is possible when the null hypothesis is assumed to be false. In these cases the empirical distribution function of p-values is not similar to the obtained from a $U([0,1])$ distribution. So in order to evaluate the performance of different distances d_i under alternative hypothesis, we

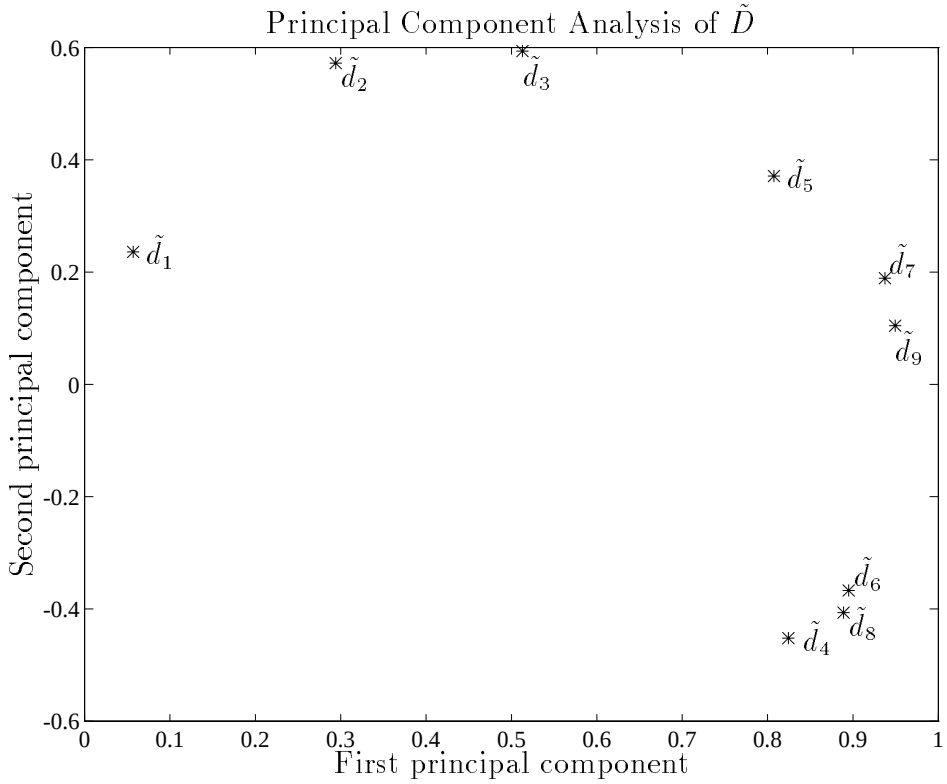

Figure 1: Principal Component Analysis of $\tilde{D}$. 10000 samples of size $n = 1000$ are generated from $U([0,1])$ and nine distances d_i are calculated between the empirical distributions and F_U.

generate data from distributions near to F_U but not exactly uniform. Distributions $\beta(a, b = 1)$, for $a = .95, .90, .85, .75$ are chosen to generate p-values, imitating the ones obtained when we separate gradually from the null hypothesis (remember that $U([0,1]) \equiv \beta(a = 1, b = 1)$). We draw 1000 samples of size $n = 500$.

Figure 2 summarizes the results of the experiment concerning the alternative hypothesis. Columns of matrix $\tilde{D}$ (calculated as before) are plotted in the plane of its two first principal components. There is a different graphic for each value of a. The graphic corresponding to H_0 (i.e., $a = 1$) is not showed, but it practically coincides with the one in Figure 1. We can see that, as data move away from H_0, global distances d_i get closer. Distances based on empirical sizes follow the other group and the farther from 1 a is, the lower remarkable the peculiarities of d_1, d_2 and d_3 are. This is especially

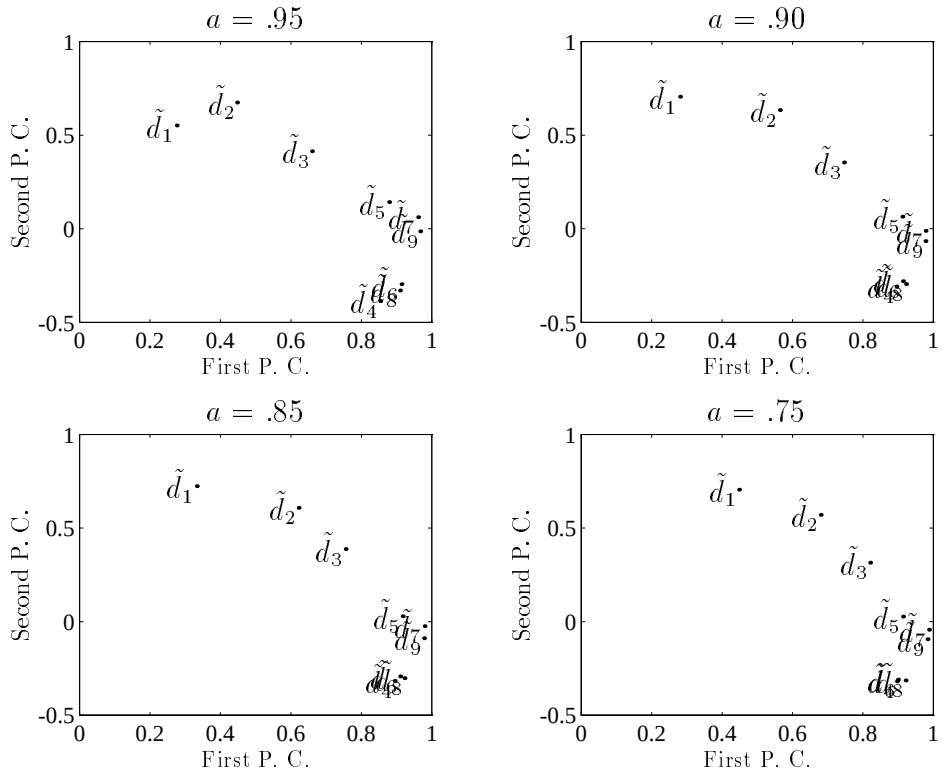

Figure 2: Principal Component Analysis of $\tilde{D}$ for nonuniform data. 1000 samples of size $n = 500$ are generated from $\beta(a, b = 1)$ and nine distances d_i are calculated between the empirical distributions and F_U.

true for d_2 and d_3. Moreover, the percentage of variance corresponding to the first principal component raises as data move away from H_0. Specifically, the first PC explains the 56%, 63%, 68%, 70% and 73% of the total variance as a equals 1, .95, .90, .85 and .75, respectively. Such result implies that the discrepancy between data and the null hypothesis (a way of interpreting the first principal component) becomes the most important feature of matrix $\tilde{D}$. As a conclusion of the analysis, the structure of interdependence between distances changes slightly when data are generated from hypothesis near H_0, and the main change is that all d_i's become closer. Therefore, differences between decisions based on different distances are more important when data are according to H_0 than when they are not.

3.2 Distribution tables of some distances

In order to make easier the use of weighted distances, we include here tables of these distances obtained by simulation of 10000 samples of uniform random variables with different sizes. Tables for unweighted distances d_4, d_6 and d_8 can be found in Shorack and Wellner (1986).

The simulated sample sizes are $n = 30$, 50, 100, 300, 500 and 1000. The empirical distribution of d_i does not change much with n, so we conclude that the asymptotic distribution of d_i is a good approximation to the distribution of d_i for every sample size n at least greater than 30. Table III shows the empirical quantile function (the inverse of the empirical distribution function) of simulated data for $n = 1000$, calculated at 100 points in $[0, 1]$. That function approximates well the asymptotic distribution of d_i.

4 Some applications

As an illustration of the performance of graphical and numerical techniques presented in this paper, we apply them to two examples. We study two hypothesis tests for which several test statistics are available and their unknown distributions can be approximated in different ways. We use the tools described above to compare all these test strategies. First, we work on the test of no correlation and second on the test of the equality of variances for two samples.

4.1 Testing the lack of correlation of two variables

Consider (X, Y) two random variables and let ρ be their correlation coefficient. We want to test $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$. Let $(x_1, y_1), ..., (x_n, y_n)$ be a random sample of (X, Y) and r_n the corresponding sample correlation coefficient. The exact distribution F_{r_n} of the statistic r_n is in general unknown.

Some tests for H_0 are based on the transformation of r_n (see, for instance, Arnold 1990, pages 419–423). The estatistic $T_n = ((n-2)^{1/2} r_n)/(1 - r_n^2)^{1/2}$ has distribution t_{n-2} when (X, Y) is a bivariate normal and $\rho = 0$. So that statistic can be used to test H_0 in the normal case. Moreover, that test procedure is asymptotically insensitive to normal assumption under fairly general assumptions. Also, the *Fisher's Z-transformation* $Z_n = .5 \log((1 +$

u	$F^{-1}_{d_5}(u)$	$F^{-1}_{d_7}(u)$	$F^{-1}_{d_9}(u)$	u	$F^{-1}_{d_5}(u)$	$F^{-1}_{d_7}(u)$	$F^{-1}_{d_9}(u)$
.00	.51288	.05795	.07711	.50	1.63556	.25509	.30636
.01	.73143	.09746	.12433	.51	1.65374	.25842	.31004
.02	.78209	.10651	.13592	.52	1.66712	.26148	.31384
.03	.84024	.11246	.14344	.53	1.67545	.26517	.31776
.04	.87732	.11767	.14958	.54	1.68898	.26864	.32153
.05	.89382	.12229	.15589	.55	1.71436	.27207	.32551
.06	.92815	.12650	.16048	.56	1.73680	.27609	.32966
.07	.95237	.13022	.16446	.57	1.75884	.27994	.33411
.08	.97240	.13360	.16866	.58	1.77447	.28434	.33825
.09	.98931	.13760	.17244	.59	1.78557	.28844	.34284
.10	1.00235	.14061	.17610	.60	1.79072	.29242	.34718
.11	1.01416	.14279	.17960	.61	1.81972	.29599	.35140
.12	1.04341	.14572	.18301	.62	1.84548	.30181	.35604
.13	1.06922	.14868	.18583	.63	1.87214	.30630	.36157
.14	1.08754	.15144	.18949	.64	1.88891	.31007	.36617
.15	1.10059	.15401	.19313	.65	1.89934	.31469	.37066
.16	1.11342	.15662	.19638	.66	1.92379	.31916	.37476
.17	1.11869	.15952	.19955	.67	1.95041	.32448	.37976
.18	1.13733	.16252	.20288	.68	1.97603	.32953	.38509
.19	1.15875	.16566	.20586	.69	1.99771	.33511	.39043
.20	1.18216	.16864	.20926	.70	2.01108	.34013	.39589
.21	1.19518	.17139	.21203	.71	2.03741	.34544	.40180
.22	1.21139	.17432	.21519	.72	2.06633	.35094	.40727
.23	1.22327	.17638	.21828	.73	2.09240	.35752	.41354
.24	1.22922	.17893	.22078	.74	2.11265	.36355	.42113
.25	1.24704	.18181	.22388	.75	2.12319	.36891	.42838
.26	1.26889	.18442	.22785	.76	2.15811	.37625	.43592
.27	1.28688	.18724	.23069	.77	2.19153	.38304	.44376
.28	1.30383	.18973	.23387	.78	2.21679	.39089	.45161
.29	1.32093	.19243	.23713	.79	2.23494	.39941	.45878
.30	1.33189	.19497	.23987	.80	2.27114	.40806	.46620
.31	1.33883	.19813	.24310	.81	2.30537	.41545	.47310
.32	1.34197	.20118	.24647	.82	2.33793	.42311	.48278
.33	1.36173	.20401	.24949	.83	2.35360	.43080	.49217
.34	1.38123	.20711	.25227	.84	2.39637	.44043	.50226
.35	1.39764	.20968	.25560	.85	2.44000	.44884	.51138
.36	1.41446	.21215	.25849	.86	2.46363	.45843	.52019
.37	1.43035	.21508	.26148	.87	2.51816	.46892	.53062
.38	1.44339	.21809	.26508	.88	2.56170	.48115	.54201
.39	1.45103	.22125	.26840	.89	2.60259	.49317	.55245
.40	1.45897	.22430	.27141	.90	2.67017	.50591	.56622
.41	1.48147	.22670	.27487	.91	2.73087	.52018	.58113
.42	1.50081	.22944	.27837	.92	2.79147	.53811	.59848
.43	1.52222	.23268	.28194	.93	2.87564	.55676	.61531
.44	1.53820	.23578	.28552	.94	2.95572	.57491	.63791
.45	1.55442	.23908	.28879	.95	3.03020	.59754	.66381
.46	1.56201	.24229	.29232	.96	3.14818	.62654	.69217
.47	1.57630	.24536	.29562	.97	3.26466	.66696	.72733
.48	1.59483	.24852	.29934	.98	3.46453	.71056	.77727
.49	1.61857	.25223	.30323	.99	3.78121	.79949	.85687

Table III: Asymptotic quantile function of d_5, d_7 and d_9 obtained by simulation. Sample size n=1000, and number of samples equal to 10000.

$r_n)/(1-r_n))$ (which is asymptotically normally distributed when $\rho = 0$) can be used to test H_0. Although both are commonly used solutions, it remains the doubt about how large must be the sample size n so that the asymptotic distribution of T_n or Z_n approximates them well.

Other test strategies, taking the sample correlation coefficient as the test statistic, may be implemented. In this paper, we study one permutation test and two bootstrap tests. All of them are based on the generation of artificial samples of the empirical correlation coefficient under H_0, that constitute a reference distribution for the statistic r_n.

In the permutation test, data y_i are shuffled and attached to data x_i to form a permuted sample $(x_i, y_i^p), i = 1, \ldots, n$, according to the null hypothesis of no correlation. Repeating B times the permutation scheme, we obtain observations of the permuted correlation coefficient r_n^*.

The first bootstrap test, based on regression, starts from the sample (x_i, y_i) and defines e_i as the residuals $(y_i - \hat{y}_i)$, corresponding to the linear regression model. Each bootstrap sample is $\{(x_i, e_i^*), i = 1, \ldots, n\}$, where e_i^* is randomly selected from $\{e_1, \ldots, e_n\}$. We draw B bootstrap samples from which B coefficients r_n^* are calculated. In Freedman (1981) some properties of that bootstrap approximation are studied.

Finally, the bootstrap test based on principal components works with the $n \times 2$ data matrix of values (x_i, y_i), $i = 1, \ldots, n$. Its principal components (PC) are calculated from the covariance matrix. The coordinates of the original pairs on the plane of PC represent a sample of n uncorrelated pairs $(\tilde{x}_i, \tilde{y}_i)$. Those pairs are resampled B times to obtain B bootstrap samples from which B coefficients r_n^* are got.

Some general conclusions are derived from simulation experiments. When data have been generated under the null hypothesis and they are near to the normality, the five mentioned ways of testing H_0 behave similarly. Nevertheless, remarkable differences appear when the distribution of data is far from the normality (for instance if they present a heavy tail). In such cases the last testing procedure (bootstrap based on principal components) differs considerably from the rest.

As an example of the obtained results, we present a simulation case. Bivariate samples of size n are generated from a mixture of two normal bivariate distributions:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \gamma_1 N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_2\right] + \gamma_2 N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 I_2\right],$$

	Permutation test			PC bootstrap test		
n	$\hat{\alpha}$	d_{KS}	$d_{L_1^w}$	$\hat{\alpha}$	d_{KS}	$d_{L_1^w}$
10	.3333	8.8335	7.5430	.3200	5.3001	4.6274
30	.3133	6.8705	5.7939	.0667	1.3799	.6308
100	.2833	5.8255	5.0662	.0633	.8603	.3579
	Critical regions at .05 significance level: $\hat{\alpha} \notin (.0253, .0747)$, $d_{KS} > 1.36$, $d_{L_1^w} > .59754$					

Table IV: Three distances for two procedures of testing $H_0 : \rho = 0$.

where $\gamma_1 = .75$, $\gamma_2 = .25$ and $\sigma = 10$. For each sample, the five above statistics are calculated and the corresponding p-values are found. When resampling techniques are used, the number of resamples is $B = 300$. We repeat the experiment $S = 300$ times for sample sizes $n = 10$, 30 and 100.

As we previously advanced, the performance of the bootstrap test based on principal components is quite different from the other four, which are very similar one another. Thus, we only report the results of two test: the later described bootstrap test and the permutation test. Figure 3 shows the empirical distribution function of the p-values for these tests when sample size ranges from 10 to 100. Table IV contains the nine distances between those empirical distribution functions and F_U. Graphical and numerical information permit to conclude that four of the five proposed test techniques show similar performance and they do not behave properly even when $n = 100$, whereas the other one (bootstrap test based on principal components) is acceptable when $n = 30$ and it works very well when $n = 100$.

4.2 Testing the equality of variances

We study now the test of the equality of variances for two samples. Several test statistics are available and their unknown distributions can be approximated in different ways. We use the tools described previously to compare all these test strategies.

Consider two random variables X and Y, with finite variances σ_X^2 and σ_Y^2. We want to test $H_0 : \sigma_X^2 = \sigma_Y^2$ against $H_1 : \sigma_X^2 < \sigma_Y^2$. We study two permutations test, one of them based on the mean and the other on the median, following the work of Baker (1995). In essence, both tests proceed as follows. Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be samples from X and Y, respectively,

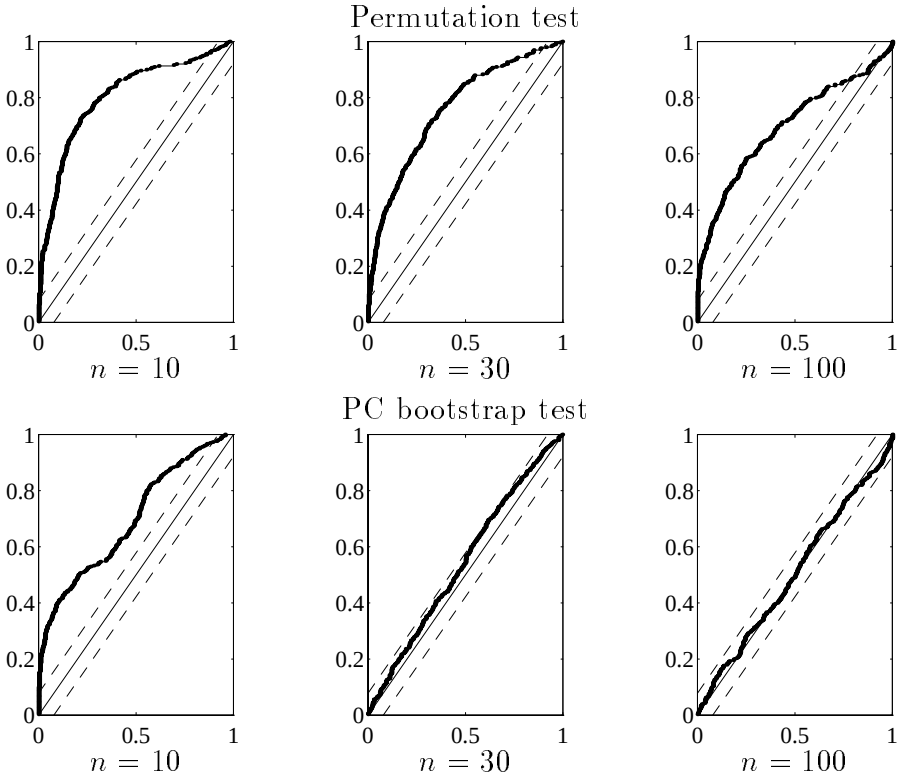

Figure 3: Empirical distribution of the p-values for two ways of testing the lack of correlation and three sample sizes.

and $\hat{x}$ and $\hat{y}$ be location estimators for those samples. Mean and median are our choices for location estimators. We define the test statistic T_n as the difference between the sums of squares $\sum_i (x_i - \hat{x})^2$ and $\sum_i (y_i - \hat{y})^2$. In order to approximate the distribution of T_n, we obtain permuted samples putting together the deviations $d_i^x = (x_i - \hat{x})$ and $d_i^y = (y_i - \hat{y})$ and randomly choosing from them two samples of size n without replacement. Then we compute the permuted version of T_n as

$$T_n^p = \sum_i (x_i^p - \hat{x}^p)^2 - \sum_i (y_i^p - \hat{y}^p)^2,$$

being x_i^p and y_i^p the elements of the permuted samples and $\hat{x}^p$ and $\hat{y}^p$ their corresponding location estimators. Repeating B times the permutation re-sampling, we have B observations of T_n^p and their corresponding empirical

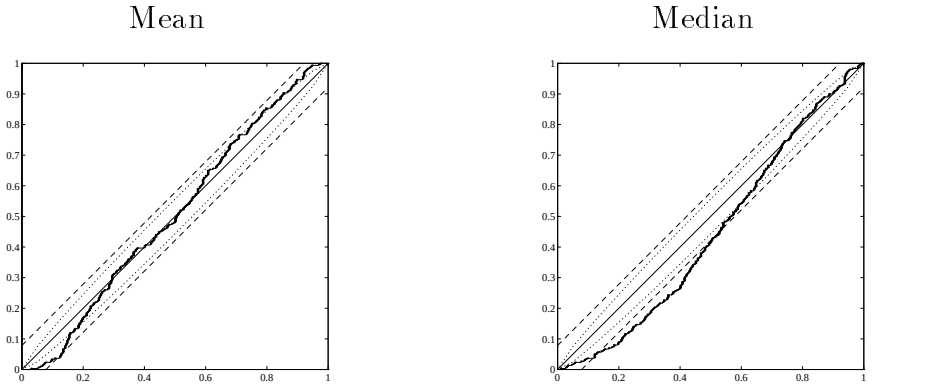

Figure 4: p-value plots for tests of the equality of variances. Dashed lines are 95% acceptance regions based on KS distance; dotted lines indicate the 95% acceptance intervals for each empirical level. Critical regions at .05 significance level: $\hat{\alpha} \notin (.025, .075)$, $d_{KS} > 1.36$, $d_{L_1^w} > .598$. Critical regions at .01 significance level: $\hat{\alpha} \notin (.018, .082)$, $d_{KS} > 1.63$, $d_{L_1^w} > .800$.

distribution is used as the approximation of the distribution of T_n.

We report graphics (Figure 4) of the empirical distributions of p-values obtained in the simulation of $S = 300$ samples of size $n = 10$. The number of permuted samples is $B = 300$. Each sample contains observations of two variables with the same double exponential distribution (thus, data are according to H_0). Empirical size for $\alpha = 0.05$ (d_2), Kolmogorov-Smirnov distance (d_4) and weighted L_1 norm (d_7) are calculated. The critical regions in the tests of uniformity of p-values based on $\hat{alpha}$, d_4 and d_7 are displayed in the caption of Figure 4. It is important to remark that conclusions about the uniformity of the p-values (and therefore about the validity of the two permutation tests) can be different depending on the distance employed. At .05 significance level, the three distances lead to reject the uniformity of the p-values distribution, but whereas d_4 and d_7 indicate that the test based on the median has worst performance than the test based on the mean, $\hat{\alpha}$ indicates the opposite. At .01 significance level, distances d_4 and d_7 indicate that only the test based on the mean is acceptable. Looking at the graphics, we visually agree with that decision.

We now pay attention to the cases where data are generated according to some alternative hypothesis, usually identified by a parameter, e.g., λ (in our

example λ is the standard deviation of Y, and a fix value for the variance of X is used). In such cases, the usual way to report results is by using graphics of the empirical power (i.e., the proportion of simulated samples for which the null hypothesis is rejected) at a given nominal size, versus the values of the parameter λ.

Our task is to find graphics, or sequences of graphics, analogous to the power function graphics. A first approach is to report a sequence of graphics that shows the empirical distribution function of the p-values for each value of λ. Figure 5 presents such a sequence of graphics when $\sigma_X = 1$ and σ_Y goes from 1 to 1.5. Two empirical distribution functions of p-values are drawn in each graphic: the one that takes higher values corresponds to the test based on the mean and the other one corresponds to the test based on the median. As we have pointed out in the introduction, to show graphically the global behaviour of a test procedure under some alternatives, requires a sequence of graphics, and it implies the use of a lot of output space. Besides, the values of λ may not be equispaced. When it occurs, we wish to be able to reflect the different size of the parameter increments in the gaps between graphics, but it does not look a trivial task.

To show the global performance of the two permutation tests when data are not according to the null hypothesis can also be done by means of distances d_i. We could compute distances from the empirical distribution function of the p-values to F_U at each value of λ, and then plot these distances as a function of λ. This is a first naive approach, because each distance d_i has its own range of acceptable values under H_0 and therefore the interpretation of those graphics is not direct. Instead of that, we could compute the probability of obtaining distances bigger than the observed ones in case that data came from the null hypothesis (remember that asymptotic distributions of distances are available; see Section 3.2). In other words, we could calculate the p-values (we call them *meta-p*-values) got from the test of uniformity of the previous p-values, based on distances d_i.

The representation of the meta-p-values versus λ provides a set of graphics with the same range of values for any distance and interpretable in the same way. The problem is now that distances d_i are usually too big as the null hypothesis is left and then meta-p-values are almost zero for alternative hypothesis not very far from H_0. Nevertheless, the calculation of meta-p-values is very interesting when data come from the null hypothesis. In the presented example, the meta-p-values of the nine distances under H_0 are dis-

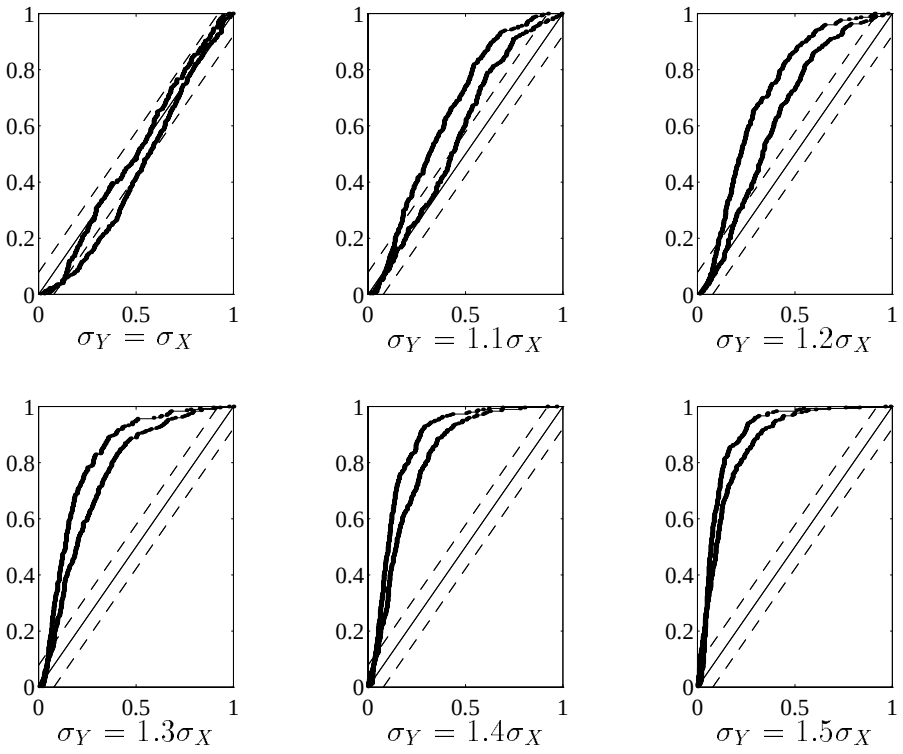

Figure 5: Empirical distribution function of p-values for data generated under some alternative hypotheses.

played in Table V. Note that the test based on the mean has an acceptable performance under H_0 if we look at global distances, and it is according to the conclusions obtained from the analysis of Figure 4.

Alternative graphics can be designed in order to achive two objectives: first, they have to be easily interpretable, independently on the involved distance, and second, graphics must not be trivial, whichever the alternative hypothesis is. Our proposal is to compute the ratio between each observed d_i and the greatest value d_i can take, which corresponds to a sequence of p-values identically equal to 1. Those ratios always lie in $[0, 1]$; they are 0 if d_i is 0 and 1 under the farthest alternative hypothesis we can suppose. Moreover, in case that d_i is d_α, this ratio is very similar to the empirical power $\hat{\alpha}$ for a nominal size α, because the ratio $|\hat{\alpha} - \alpha|/(1 - \alpha)$ is approximately $\hat{\alpha}$ when

Distances	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
Mean	.0741	.0002	.0000	.0181	.0001	.0842	.0352	.0754	.0222
Median	.0741	.0081	.0002	.0000	.0000	.0003	.0000	.0001	.0000

Table V: Meta-p-values for the permutation tests based on the mean and on the median in the test of the equality of variances.

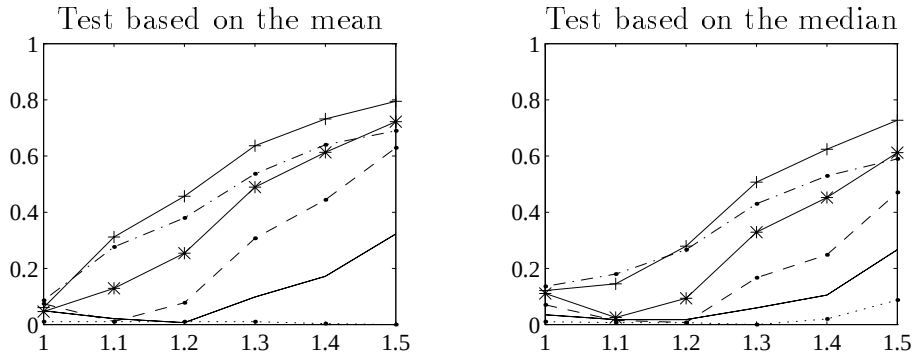

Figure 6: Ratios of observed distances d_i over its supremum values in the test of the equality of variances. Doted line corresponds to $d_{.01}$, solid line to $d_{.05}$, dashed line to $d_{.1}$, dashdot line to d_{KS}, plus symbols to L_1 and star symbols to $d_{L_1^w}$.

α is small. For d_{KS}, the ratio is again d_{KS}. The supremum of distances d_{L_1} and d_{L_2} are .5 and $\sqrt{1/3}$, respectively. In case of d_{KS}^w, the supremum is $\sup_{u\in[0,1]}(1-u)w(u)$, and it is equal to 3.411 for the function w used in that paper. The value of supremums for the weighted L_1 and L_2 norms are, respectively, $(1-\mu_w)$ and $\sigma_w^2+(1-\mu_w)^2$, where μ_w is the expectation of a random variable with density function w and σ_w^2 is its variance. For the weight function w used in this work, these two supremum are 0.8 and .7855.

In Figure 6, we plot ratios of the observed distances d_i over its supremum values. As weighted distances show very similar ratio patterns, we only plot the ratio corresponding to L_1^w. For the same reason, we plot the L_1 ratio instead of plotting the ones for both L_1 and L_2. In the test based on the median, the passing of the empirical function of p-values from the lower diagonal region into the upper diagonal region is also reflected in the ratio graphics of weighted distances: these ratios have a minimum not at $\sigma_Y=\sigma_X$ but at $\sigma_Y=1.1\sigma_X$. Finally, we can conclude that the test based on the mean is better than the test based on the median, because the ratios corresponding

to the mean are always higher than the ones corresponding to the median. The implications of this fact are similar to those derived from the observation of two power functions, if one of them was always over the other one. The same conclusion is also derived from Figure 5.

5 Conclusions

In this paper, we have presented new graphical and numerical tools that summarize the results of a simulation study concerning hypothesis test. These tools are mainly based on the computation of distances between distribution functions. We have tabulated the null distribution of some of these distances. A joint study of several distances reveals that important aspects of a test can pass unnoticed if only empirical significance levels are calculated. The proposed tools have been applied in two practical examples, demonstrating their usefulness in the discrimination between alternative test procedures and also in the detection of data not according to the null hypothesis.

References

Arnold, S. F. (1990). *Mathematical Statistics*. Prentice-Hall.

Baker, R. D. (1995). Two permutation tests of equality of variances. *Statistics and Computing*, **5**, 289–296.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey (1983). *Graphical Methods for Data Analysis*. Wadsworth and Brooks.

Davidson, R. and J. G. MacKinnon (1994). Graphical methods for investigating the size and power of hypothesis tests. Discussion Paper 903, Queen's Institute for Economic Research. Revised March, 1997 (http://qed.econ.queensu.ca/pub/faculty/mackinnon/).

Delicado, P. (1995). *Contrastes de bondad de ajuste en el modelo de regresión con coeficientes aleatorios*. Ph. D. thesis, Univ. Carlos III de Madrid.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, **9**, 1218–1228.

Shorack, G. R. and J. A: Wellner (1986). *Empirical Processes with Applications to Statistics*. Wiley.